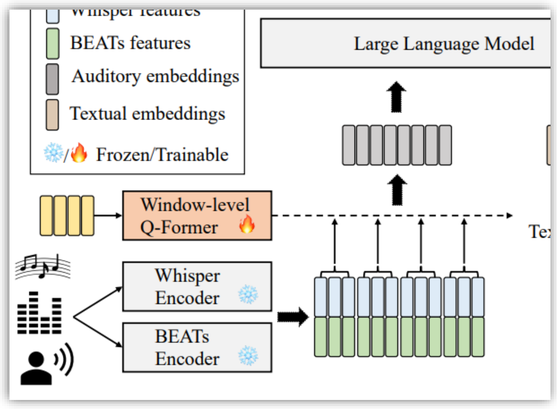
Enhancing Speech Large Language Models with Prompt-Aware Mixture of Audio Encoders

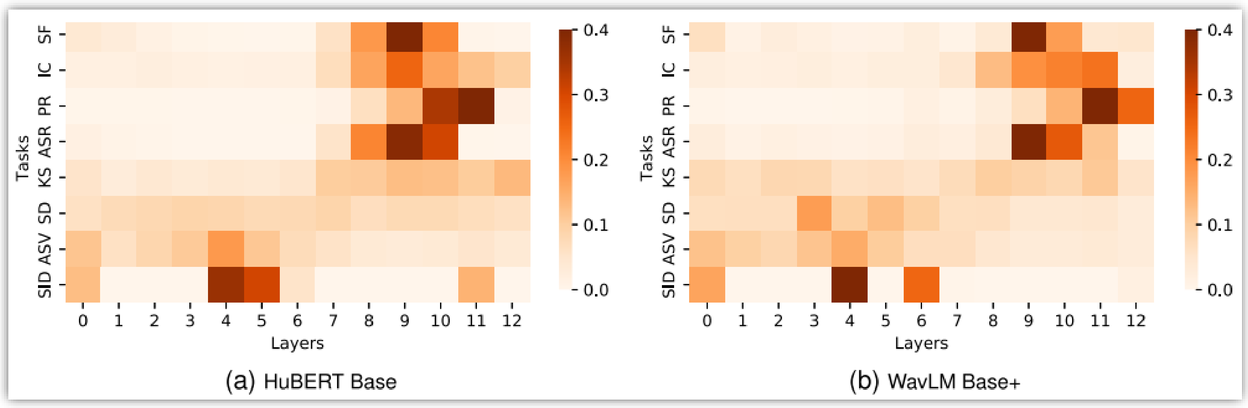
Weiqiao Shan[1], Yuang Li[2], Yuhao Zhang[3], Yingfeng Luo[1], Chen Xu[4], Xiaofeng Zhao[2],
Long Meng[1], Yunfei Lu[2], Min Zhang[2], Hao Yang[2], Tong Xiao[1,5†], Jingbo Zhu[1,5]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]Huawei Translation Services Center, Beijing, China
[3]The Chinese University of Hong Kong, Shenzhen, China
[4]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[5]NiuTrans Research, Shenyang, China

Motivation

1. More and more speech LLM have been adopting dual-encoder designs.
2. Distinct feature requirements of different audio downstream tasks are manifested across different layers.

The **Whisper** … **suitable to model speech and** … **background noises**. **BEATs** is trained to **extract** … **semantics information** …





- Insight 1: **Single** audio **encoder cannot** sufficiently capture all the features demanded by **multiple downstream tasks**.
- Insight 2: **Features** extracted **from different layers** of the audio encoder contribute with **varying importance to different tasks**.

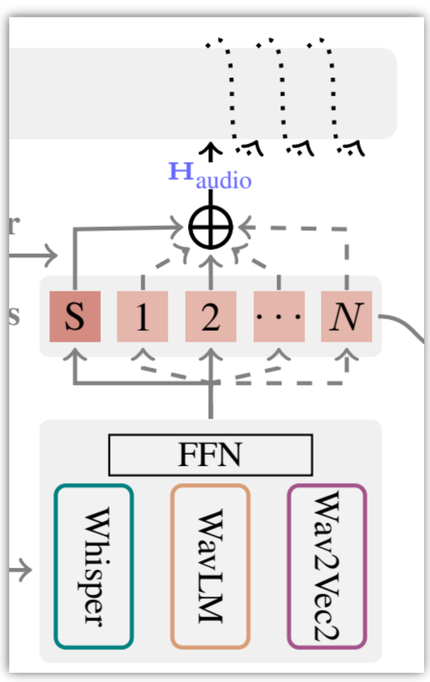
Prompt-aware Mixture (PaM)

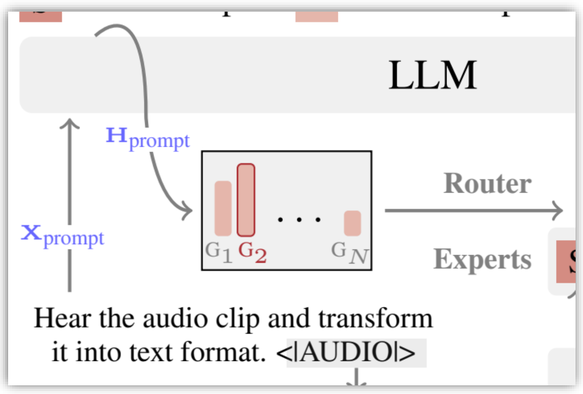
Task-specific information in speech processing tasks is typically embedded in prompts

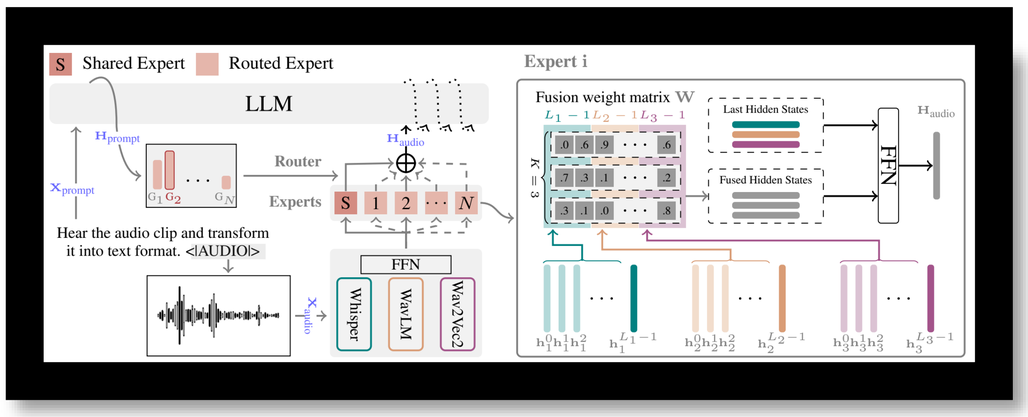
1. Extracting the hidden representations from each layer of multiple encoders and mapping them into a unified dimension.

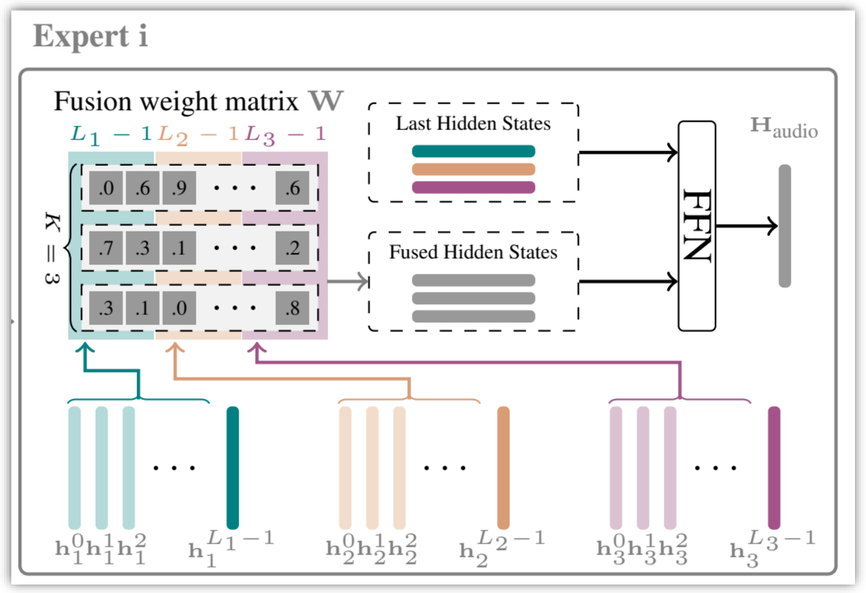
2.Encode the input prompt using LLMs, obtain its last hidden states, and map it into a routing weight.

3. In each expert, We use *K* sets of distinct fusion parameters to integrate the hidden states from all layers of all encoders (excluding the last hidden state of each encoder), and yielding *K* fused hidden states.





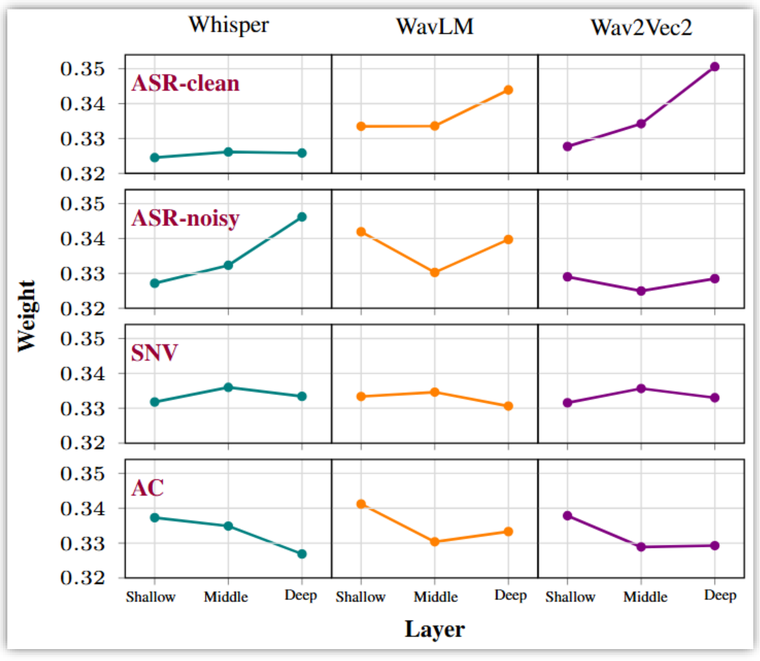




Results

PaM surpasses both single-encoder and multi-encoder baselines across a variety of tasks and datasets

Model	LibriSpeech		AMI	SNV	AudioCaps	AudioCaps QA	AVG Rank↓
	WER(clean)↓	WER(other)↓	WER↓	Acc↑	METEOR↑	METEOR↑	
Single-encoder Baselines							
- Whisper (Radford et al., 2022)	9.61	16.73	**16.27**	18.80%	**32.96**	**15.04**	5.67
- WavLM (Chen et al., 2022)	5.59	10.57	18.97	**41.40%**	27.14	12.77	5.50
- Wav2Vec2 (Baevski et al., 2020)	**4.30**	**09.46**	26.69	39.00%	23.81	11.20	5.33
Multi-encoder Baselines							
- WavLLM (Hu et al., 2024)	4.95 (- 0.65)	09.19 (+0.27)	15.29 (+0.98)	39.20% (- 2.20)	34.93 (+1.97)	16.35 (+1.31)	2.67
- SALMONN (Tang et al., 2024)	5.04 (- 0.74)	09.70 (- 0.24)	19.04 (- 2.77)	49.40% (+8.00)	34.86 (+1.90)	15.97 (+0.93)	3.50
- Average	4.76 (- 0.46)	10.43 (- 0.97)	17.20 (- 0.93)	45.50% (+4.10)	33.22 (+0.26)	15.53 (+0.49)	3.67
PaM (Ours)	3.65 (+0.65)	07.07 (+2.39)	12.79 (+3.48)	42.80% (+1.40)	35.47 (+2.51)	15.70 (+0.66)	**1.67**



Encoders	LibriSpeech		AMI	SNV	AudioCaps	AudioCaps QA	AVG Rank↓
	WER(clean)↓	WER(other)↓	WER↓	Acc↑	METEOR↑	METEOR↑	
Whisper+X	4.42	8.92	13.96	43.70%	**35.41**	16.11	4.5
WavLM+X	4.07	7.97	18.47	47.47%	32.48	15.01	5.5
Wav2Vec2+X	**3.42**	7.20	18.18	45.13%	32.33	15.02	4.3
HuBERT+X	3.87	8.21	19.49	36.90%	31.82	15.08	6.8
Whisper+X+Y	4.22	8.11	**13.79**	**49.77%**	35.37	**16.31**	**3.0**
WavLM+X+Y	3.72	7.99	16.35	38.73%	34.23	15.82	4.0
Wav2Vec2+X+Y	3.77	**6.90**	16.90	42.63%	34.23	15.82	3.8
HuBERT+X+Y	4.03	7.96	17.13	44.17%	34.18	16.13	4.0

- Different downstream tasks depend on different encoder layers.
- Employing more encoders enhances the performance of the speech LLM on all downstream tasks.

Paper

Code